

DOCUMENT RESUME

ED 148 903

TH 006 897

AUTHOR Haladyna, Tom; Roid, Gale
TITLE A Theoretical and Empirical Comparison of Three Approaches to Achievement Testing.
PUB DATE [76]
NOTE 26p.
EDRS PRICE MF-\$0.83 HC-\$2.06 Plus Postage.
DESCRIPTORS Academic Achievement; *Achievement Tests; *Comparative Analysis; Criterion Referenced Tests; Dentistry; Educational Objectives; Evaluation Criteria; Higher Education; Item Analysis; Mathematical Models; Measurement Techniques; Norm Referenced Tests; *Sampling; Standard Error of Measurement; *Statistical Analysis; *Test Construction; Test Interpretation; Test Items; Test Reliability; Test Validity; *Theories
IDENTIFIERS Domain Referenced Tests; Latent Trait Models; *Test Theory

ABSTRACT

Three approaches to the construction of achievement tests are compared: construct, operational, and empirical. The construct approach is based upon classical test theory and measures an abstract representation of the instructional objectives. The operational approach specifies instructional intent through instructional objectives, facet design, item forms, and item-generating algorithms. Criterion referenced and domain referenced tests are discussed, as they relate to the operational approach. A series of trials and subsequent calibrations are used in the empirical approach, which is based upon latent trait theory. To compare these three approaches, an achievement test was administered to 258 dental students, before and after instruction in basic dental anatomy. Five different strategies were used to construct the test forms: random selection, item-test correlations, pre-post test difficulty index, Bayesian, and Z-difference. Descriptive statistics were computed, including estimated and actual reliability, and estimated and true standard errors of measurement. The results suggested that a random sampling approach provides the greatest representativeness and minimizes errors of measurement. Random sampling is recommended for both the construct and operational definition approaches to measurement. A completely valid test of the third approach has not yet been completed. (Author/GDC)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

A THEORETICAL AND EMPIRICAL COMPARISON OF THREE APPROACHES TO ACHIEVEMENT TESTING¹

Tom Haladyna and Gale Roid
Teaching Research Division
Oregon State System of Higher Education

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

T. HALADYNA

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC) AND
USERS OF THE ERIC SYSTEM."

ED148903

Classroom achievement testing has typically involved the selection of test items representing content areas or important aspects of what is to be taught. In measurement textbooks, readers are frequently encouraged to select test items based on the relationships between the item and test performance for a group of examinees. The advent of many forms of systematic instruction, e.g., mastery learning (Bloom, 1968), individually prescribed instruction (Glaser, 1967), and the personalized system of instruction (Keller, 1968), as well as criterion-referenced evaluation of test results has led measurement specialists to doubt the veracity of traditional practices and to explore new conceptualizations. As a consequence there have been a number of interesting and promising approaches to item construction, studying item characteristics, estimating reliability and improving the validity of classroom achievement tests.

It is clear that in any form of systematic instruction there is a need for an explicit relationship between the goals (or objectives) of instruction and the test upon which student achievement is evaluated. This point cannot be underemphasized. Shoemaker (1975, p. 144) states that:

...the item universe associated with an instructional program is a critical component of the program and it is, indeed, its raison d'être.

Thus, it is clear that the identification or creation of groups of items

¹This paper represents work in progress. Critical comments or suggestions are invited by the authors. No reference should be made to this paper without consent of the authors.

representing the content each student needs to learn is a necessary process in the planning of instruction. As Shoemaker points out, the idea of item universes is not new, but its application to instructional programs and achievement tests is. The procedures one should follow in constructing such tests raise important theoretical and empirical issues.

PROBLEM

There are at least three theoretical schemas for constructing classroom achievement measures: (a) construct, (b) operational definition, and (c) empirically-based approaches.

The Construct Approach

According to Cronbach and Meehl (1955, p. 283) a construct is "...some postulated attribute of people, assumed to be reflected in test performance." To study the validity of the responses of persons to a measure of a construct one needs to (a) formulate a construct in purely abstract terms, (b) explicate that construct in terms of observables, and (c) marshal empirical evidence to support the existence of the construct. Examples of empirical evidence include group differences, correlations between two or more tests designed to measure the construct, factor analysis of test items related to the construct, or studies of predicted change over occasions. Each of the types of evidence must be consistent with the formulation of the construct and predictions based on the theoretical network of relationships between the observables associated with the construct. The very essence of a measure of a construct is in its interpretation where test performance is linked to an abstract entity.

In the realm of achievement testing, a construct can be any instructionally relevant trait. One describes these constructs in terms of a topic, notes, a text, an outline, instructional objectives, overviews, or advanced organizers. According to the tenets of classical test theory, one can measure a construct

by first specifying an item universe and then randomly sampling items from that universe to create test forms. In practice, item universes were seldom available, and writers of test and measurement texts have long advocated the building of classroom achievement tests through logical and empirical processes which yield high content validity. One common practice has been to assemble a test composed of items judged to measure the construct and then to refine the test by removing items that are not highly discriminating. In fact, most writers of testing textbooks advocate the building of tests by selecting items with the highest item-test correlation (discrimination index). One rationale for this practice is that the total test score is a global measure of the construct and items related to total scores are the better indicators of the construct. Popham and Husek (1969) have maintained that point-biserial correlation indexes are subject to invalidity because the sample of examinees is restricted with respect to achievement and the correlations are attenuated. Haladyna (1974) used unrestricted samples consisting of instructed and uninstructed students which made such correlations more adequately estimable. Several research studies have shown that these discrimination indexes are not similar to those constructed using posttest only samples (Haladyna and Roid, Note 1; Helmstadter, Note 2).

In conclusion, although the construct approach to classroom achievement testing is theoretically based on defining item populations and random sampling, practical methodology has involved the use of empirical methods for item selection coupled with logical analysis of what content should be tested. The need for a more rigorous construct approach has recently been championed by Messick (1975).

Operational Approach

An operational definition describes in terms of observables a behavior or set of behaviors that are to be measured. If the operational definition

is acceptable to all those employing measures derived from the definition, then the measure is directly interpretable and constitutes a criterion (Cronbach and Meehl, 1955).

Following this approach to its logical end, a test can be designed that is simply a random sample of the set of tasks that are operationally defined. As Osburn (1968) has argued, we are interested in generalizing from a set of items on a test to a larger universe of items, and an operational definition of the universe of items provides the basis for such a generalization. The interpretation of test performance is based on the meaning assigned to that set of tasks, and no attempt is made to infer meaning to something abstract.

In classroom achievement testing, the use of instructional objectives and later the development of item writing theories (e.g., Bormuth's linguistic analysis, 1970; Guttman's facet analysis, 1969; item forms, Hively, Patterson and Page, 1969) represent an application of the operational definition approach to testing. Each of the above-named, item-generating procedures leads to specifiable item universes from which achievement tests can be constructed. When tests are constructed to measure a single instructional objective or set of objectives, they are often termed "criterion-referenced," while the term "domain-referenced" is used to denote the usage of an item-generating algorithm and the random sampling of items to test forms. Hively, et al. (1969) characterized the operational definition approach to testing as educational "test behaviorism."

It is clear from the literature related to the construction of criterion-referenced tests, that two test construction practices are most often recommended (e.g., Crehan, 1974; Haladyna and Roid, Note 1; Kosecoff and Klein, Note 3): (a) empirical item selection or (b) random sampling of items. The first represents the analogue to empirical procedures with the construct approach, that of selecting items that show the greatest differences in item

difficulties computed from uninstructed and instructed samples. The item indexes used can be termed "instructional sensitivity indexes," because they identify items that are sensitive to the changes that occur in the test performance of learners before and after systematic instruction.

A plethora of item sensitivity indexes have been recommended for tests representing the operational definition approach. These include the pre-post difficulty index (PPDI) suggested by Cox and Vargas (Note 4), a Bayesian index (Helmstadter, Note 5), and a Z-difference statistic based on a one-parameter latent trait model, the Rasch model (Haladyna and Roid, Note 1). Each of these were found by Haladyna and Roid to have a high degree of similarity.

In domain-referenced test construction, the item generation procedure is used to develop an item universe, and then items are randomly selected for each test form. The interpretation that follows is that consistent test responses provide an unbiased estimate of performance on the entire item universe (Osburn, 1968).

Empirical Determination

One may study a trait simply by identifying groups of items which are believed to measure a trait and seek empirical confirmation of this hypothesis. This approach requires a statistical rationale rather than the theoretical premise which exists in the first two approaches. To empirically determine which items are and are not measures of a trait, one may employ one of several latent trait models (Lord and Novick, 1968).

In latent trait theory, examinee ability is believed to be related to performance on each test item with respect to a logistic item characteristic curve. The application of the theory leads to an item-free estimate of student achievement and sample free estimates of item difficulty. A mean-square-fit

statistic is used to determine if an item does or does not fit the latent trait. Items that pass this statistical test are considered appropriate for measuring the latent trait. Items are then "calibrated" by determining their difficulty, and in the more complex models, by determining difficulty, discrimination and a third "guessing" parameter. A calibrated item universe is created for the purpose of testing examinees.

Latent trait theory may be reasonably applied to achievement testing in a number of ways. One of the particular advantages of a latent trait approach is that tests can be constructed for an examinee which will minimize the error of measurement. Therefore, items are selected on the basis of their calibrations because errors of measurement are believed to be smallest when each examinee is matched with items from the calibrated item pool which are of an appropriate difficulty.

To illustrate, if a student is anticipated to be a higher achiever (that is, the student's true level of attainment is 80% of the items in the universe), errors of measurement are smallest if items are chosen from the domain so that the student would miss about 50% of the items. Regardless of the particular set of items administered, students' scores can be computed that are based on a metric that is independent of the immediate test form. One strategy that uses latent trait theory, is to increase the precision of measurement by giving students items which are appropriate to their level of attainment.

Comparison of the Three Approaches

The characteristics of these three approaches to test construction are given in Table 1. The interpretation one wants to make about a test score seems to determine what approach, construct or operational definition, is required. The procedures for constructing a test representative of either theoretical approach is random sampling. The basic difference in the two theories is in how test results are interpreted. Despite this insistence

Table 1

	Construct Approach	Operational Definition	Empirical Determination
Common Name	Norm-Referenced	Domain-Referenced or Criterion-Referenced	Latent Trait
Item Selection or Generating Procedure	(1) abstractly define the trait (2) intuitively and subject- ively write or select items	(1) Item writing algorithms (2) instructional objectives	field test a collection of items and calibrate
Test Construction Strategy	(1) randomly sample from uni- verse of items or (2) select items with highest item-test correlation	(1) randomly sample from the universe of items or (2) select items with the highest instructional sensitivity index	select items for test that are most appropriate with respect to difficulty to the level of attainment of the examinee
Empirical Criteria for Judging Test Quality	(1) KR-20 estimate of reliability (2) item-test correlations (discrimination indexes) (3) inter-item correlation (internal consistency)	(1) a wide variety of new reliability estimates (2) instructional sensitivity indexes (3) differences in means of instructed and uninstruc- ted examinees	standard error of estimate for a particular examinee

on random sampling from a well defined item universe, applications have involved both logical and empirical procedures for test construction. In the case of empirical determination with the latent trait method, the selection of items to "fit" the examinee represents an important alternative test construction methodology.

Does one choose test items from an item universe based on empirical criteria or randomly select items? What are the consequences of these alternative procedures on item and test characteristics and more importantly, on student scores and their associated errors of measurement?

Research Questions

To answer some of the questions above, a set of definitions are needed regarding a general theory of measurement and specific research questions need to be posed. Definitions appear in Table 2 and are worthy of consideration, as they are crucial to the concepts previously discussed and necessary for the empirical aspects of this study.

The underlying assumptions and procedures of the present research are as follow. A group of test items that represent an achievement domain can be identified using a specified procedure, e.g., an operational definition approach. True scores can be obtained by administering the universe of items to all examinees, both before and after instruction. The difference between pre-instruction and post-instruction means is a parametric measure of the effects of instruction. Scores of all examinees on the universe of items constitute their true scores.

Different test forms may be created by selecting items with high item-test correlations or high instructional sensitivity indexes or items can be randomly selected. These test forms, when scored for the same examinee, provide observed (fallible) scores which are estimates of true

Table 2

AVERAGE ABSOLUTE DIFFERENCE (AAD):

$$AAD = \frac{\sum |X_t - X_e|}{n}$$

ESTIMATES OF STANDARD ERRORS OF MEASUREMENT: obtained by

$$\hat{\sigma}_{SEM} = s.d. \sqrt{1 - r_H}$$

where r_H is reliability estimate

OBSERVED SCORE: the score obtained by administering any set of items of the universe to any student from the population of students for which instruction is intended.

RELIABILITY: the squared correlations between true and observed scores (r_H)

RELIABILITY ESTIMATES: a statistical procedure whereby test results are employed to estimate the proportion of total test variance which is true variance (r_H).

STANDARD ERROR OF MEASUREMENT: the standard deviation of errors of measurement. Obtained by:

$$\sigma_e = \sqrt{\frac{\sum_{i=1}^n (X_{ti} - X_{oi})^2}{n}}$$

where X_t is a true score
 X_o is an observed score
 n is the number of examinees

TRUE SCORE: the score obtained by administering all items of the universe to any student from the populations of students for which instruction is intended.

UNIVERSE: a large population of items representing some instructional intent.

UNIVERSE SCORES: the score received by any student from the population of interest based on the administration of every item from that universe of items.

scores.² The degree to which these observed scores differ from the true scores constitute errors of measurement, and the squared correlation between true and observed scores for each form is the reliability. Both reliability and errors of measurement can be estimated using the conventional KR-20 formula and the formula for the standard error of measurement shown in Table 2. A statistic useful for examining the magnitude of error for any test when true scores are known, which was suggested by Hambleton, Hutton and Swaminathan (1976), is the average absolute difference (AAD). The formula is given in Table 2.

As a consequence of the theoretical analyses of the problem and the research reported in this section, the following specific research questions were investigated:

1. How do the means and standard deviations of the test forms constructed using different item selection paradigms compare with the parametric means and parametric standard deviations across the dimensions of different test lengths and different sample conditions?
2. How accurate are the conventional estimates of reliability and standard errors of measurement with respect to true reliability standard error of measurement for these various test forms and testing conditions?

METHOD

Procedure

A 100-item achievement test was administered before and after instruction to 258 dental students. The achievement domain was basic dental anatomy represented by a five-volume programmed text on the topic. By virtue of the

² In classical test theory the observed score is said to be biased. In domain-referenced testing it is unbiased. See Nunnally (1967) and Millman (1974b) for a discussion of these definitions.

procedures employed to construct this 100-item test, it can be considered relevant to the objectives of the text. The test may be considered as a measure of a construct, e.g., "knowledge of basic dental anatomy," or as a collection of items operationally defined by the objectives of the texts, depending upon one's perception of the intent of instruction that employs the texts. Several important assumptions were made for the purpose of conducting this study and for the relevance of the theoretical positions represented. Achievement domains are believed to be specifiable in terms of universes of items. The 100-items were identified as a universe of items for which achievement tests would be created and administered to students. Test forms were constructed consisting of 20 and 50 items each, representing five different test construction strategies: (a) random selection, (b) item-test correlations, (c) pre-post difficulty index, PPD1, (d) Bayesian, and (e) Z-difference. It was decided that data would be analyzed with two types of samples, a post-instruction sample representing instructed students and a combined pre-and-post instruction sample representing the full range of levels of attainment in the achievement domain.

Analyses

Descriptive statistics were computed for all students on each test form and the 100-item universe across the sample conditions. Correlations were computed and scattergrams were done for relationships between scores from each test form and 100-item scores for the restricted and non-restricted samples. The average absolute difference (AAD) was computed for all student scores for each form. Conventional KR-20 estimates and the true reliability were computed for each test form, as well as estimated and true standard errors of measurement.

RESULTS AND DISCUSSION

Two criteria were employed to evaluate the alternative test construction strategies: the relationships between (a) observed and parametric (true) means and standard deviations and (b) observed and parametric reliabilities and standard errors of measurement.

Comparison of Observed and Parametric Means and Standard Deviations

The means and standard deviations of all test forms for the 258 examinees using the restricted sample and the unrestricted samples appear in Table 3. The parametric mean and standard deviation for the restricted sample were 74.5% and 7.9%. For the unrestricted sample the mean and standard deviation were 53.4% and 23.6%. Comparing these various test forms within respective sample types, the random sample test construction strategy yielded test forms that provided the closest approximations to the parametric means under all circumstances. The Z-difference statistic led to the construction of forms that were quite close to the parametric mean in the case of the 50-item length, based on restricted or unrestricted samples.

Standard deviations were closest to the parametric standard deviation for the four 20-item empirically-determined test forms for the restricted sample. In the other three instances, the randomly assembled test forms provided the closest estimates to the parametric standard deviation. The Z-difference test form for the unrestricted 50-item test proved to be remarkably close to the parametric standard deviation.

Simple F-tests were done to examine the differences in variance between each of these test forms' variances and the parametric variance. The results indicated that when using the restricted sample, 20-items, randomly-constructed test forms had statistically significant ($p < .05$) higher variances than the parametric variance. For the unrestricted sample, the opposite was true, the

Table 3

Means¹ and Standard Deviations¹ of Test Forms Constructed Using Different
(a) Test Construction Strategies, (b) Test Lengths, and (c) Samples

Test Construction Strategy	Restricted ² 20-item		Unrestricted ³ 20-item		Restricted ² 50-item		Unrestricted ³ 50-item	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
1. Random	77.9	9.7	56.9	24.4	73.4	8.1	53.6	22.9
2. Random	67.8	13.3	48.6	23.7	71.7	8.5	51.8	22.8
3. Random	76.9	11.1	52.8	27.6	71.1	8.8	50.6	23.3
4. Point Biserial	91.1	8.6	60.3	34.4	85.1	8.2	58.6	29.3
5. Bayesian	92.4	7.9	63.8	32.3	87.6	7.4	62.6	28.1
6. PPD1	87.8	9.6	56.2	34.5	82.6	8.6	55.9	29.4
7. Z-difference	90.4	8.9	59.4	34.3	74.2	8.8	53.6	23.8
8. Domain (true scores)	74.5	7.9	53.4	23.6	74.5	7.9	53.4	23.6

1. All test scores are reported in the metric of percentage correct.

2. Post-instruction examinees only.

3. Pre- and post-instruction examinees.

15/16

four empirical test construction techniques for the 20-item test forms yielded higher variances. With the 50-item test forms there were no significant differences.

Comparison of Estimated and True Reliabilities and Standard Errors of Measurement

The distinction was made earlier between reliability and reliability estimates as well as the true standard error of measurement and the estimated standard error of measurement. The present study investigated the effect of various test construction strategies on the actual reliabilities and errors of measurement in contrast to estimated reliabilities and standard errors. The results are presented in Table 4 for the 20- and 50-item forms for the restricted and unrestricted samples.

When examining the reliabilities across the four conditions of the study, random selection of items led to test forms which were usually more reliable than empirically-based tests with a restricted sample. With unrestricted samples, reliabilities were remarkably similar despite the fact that the variances of test scores from various test forms varied considerably.

Reliability estimates were lower than corresponding true reliabilities in 24 of 28 instances. The four exceptions were empirically-constructed test forms under the condition of an unrestricted sample. Each of these were extremely close to true reliabilities. Overall, the use of unrestricted samples led to more accurate reliability estimates.

Errors of measurement were smallest with test forms created by random sampling of items in three of four conditions, 20-item forms with restricted samples proved to be the lone exception. These results would suggest that randomly-constructed test forms provide the best estimate of true scores under most conditions.

Table 4

Reliability¹, Reliability Estimates² and Standard Errors of Measurement,
Estimated Standard Errors of Measurement for Test Forms Which are Based
On Different (a) Test Construction Strategies, (b) Test Lengths, and
(c) Samples

Test Construction Strategy	Restricted 20-item				Unrestricted 20-item				Restricted 50-item				Unrestricted 50-item			
	Rel.	Rel. Est.	SEM	Est. SEM	Rel.	Rel. Est.	SEM	Est. SEM	Rel.	Rel. Est.	SEM	Est. SEM	Rel.	Rel. Est.	SEM	Est. SEM
1. Random	.534	.280	6.6	8.2	.913	.850	7.2	9.4	.793	.549	3.7	5.4	.967	.932	4.2	6.0
2. Random	.645	.527	7.9	9.1	.861	.827	8.8	9.8	.810	.586	3.7	5.5	.971	.931	3.9	6.0
3. Random	.556	.418	7.4	8.5	.916	.889	8.0	9.2	.806	.605	3.9	5.5	.975	.936	4.0	5.9
4. Point Biserial	.405	.292	6.6	7.2	.928	.932	9.2	9.0	.782	.651	3.8	4.8	.973	.965	4.8	5.5
5. Bayesian	.344	.226	6.4	7.0	.921	.924	9.1	8.9	.696	.614	4.1	4.6	.968	.963	5.0	5.4
6. PPD1	.468	.399	7.0	7.4	.923	.935	9.6	8.8	.784	.658	4.0	5.0	.972	.964	4.9	5.6
7. Z-difference	.428	.334	6.7	7.3	.929	.931	9.1	9.0	.792	.613	4.0	5.5	.968	.939	4.3	5.9

1. Squared correlation of true and observed scores.

2. Based on KR-20.

When reliability is underestimated, errors of measurement are overestimated. The degree of error in estimating can be found in comparing actual standard errors of measurement with estimated, shown in Table 4. While none of these differences are statistically significant, as determined by simple F-tests, the magnitudes of these errors, as they consistently occur, yields information about the reasonableness of using a conventional reliability analysis followed by the application of the formula for the standard error of measurement. The only condition where estimated standard errors of measurement approached actual standard errors was with the empirically-constructed 20-item test forms using an unrestricted sample. No single item selection strategy provided the most reliable test form or accurate standard error of measurement.

An index of measurement precision, the average absolute difference (AAD) for each examinee was computed for each test form and is shown in Table 5. Regardless of the sample type or test length, random sampling of items led to the creation of the most precise test forms. AAD's were smallest for all random forms when compared to empirically-determined forms. One exception to this result was observed for the 50-item test form constructed using the Z-difference statistic. In both the restricted and unrestricted samples, the Z-difference test had a low AAD. Sample type did not seem to influence the magnitudes of AAD's.

CONCLUSIONS

1. If one is interested in producing tests which approximate an item universe with respect to means and standard deviations, random sampling of items from that domain is probably the most effective. In fact, greater precision is anticipated if one uses a stratified sampling plan as suggested by Millman (1974a).

Table 5

	AAD	20-item Range of Errors	AAD	50-item Range of Errors
Restricted Sample				
1. Random	6.39	0-23	3.32	0-14
2. Random	7.97	0-34	3.44	0-12
3. Random	6.73	0-25	4.00	0-15
4. Point Biserial	12.78	0-34	7.32	0-19
5. Bayesian	12.93	0-39	9.57	0-26
6. PPD1	12.07	0-33	6.39	0-21
7. Z-difference	12.37	0-36	3.39	0-15
Unrestricted Sample				
1. Random	6.00	0-21	3.09	0-11
2. Random	8.46	0-34	3.79	0-12
3. Random	6.09	0-24	4.19	0-15
4. Point Biserial	16.71	0-34	10.57	2-19
5. Bayesian	17.92	0-39	13.07	2-26
6. PPD1	13.67	0-33	8.18	0-21
7. Z-difference	15.97	0-36	3.24	0-13

2. With respect to properly estimating the parameters of mean and standard deviation, the restricted sample (posttest only) appears to have some limitations which obscure true relationships.

3. If there is a concern to produce accurate estimates of errors of measurement, no procedure or condition employed in this study appears to provide accurate estimates. Using a conventional KR-20 procedure does not lead to accurate estimates of reliability and the problem is heightened when the standard error of measurement is calculated. Reliability is typically underestimated while standard errors are consistently overestimated.

4. Tests which are created by random sampling seem to provide the smallest errors of measurement as well as average absolute differences between true and observed scores. If the concern is estimating true scores, as defined in this study, the scores obtained from tests where items are randomly sampled appear to be the most useful.

5. The results of this study firmly support the notion that tests ought to be constructed by the random sampling of items to test forms, if observed means, standard deviations, and examinee scores are to be used to estimate parametric values, as defined in this study.

This should not eliminate the need to empirically study test items. In fact, the creation of item universes is based on an emerging technology which does not guarantee high quality items (Roid and Haladyna, in press). Consequently, continued attempts at refining a general theory for achievement testing should be concerned with empirical screening of test items or item forms prior to entry into item pools for test construction. This procedure should not conflict with construct or operational definition interpretations of achievement tests.

The third approach to achievement testing, the empirical determination of measures of latent traits, was reviewed but not empirically investigated

in this study. This work needs to be done and compared with results obtained in studies like the present one. Empirical screening of items in latent trait theory appears to be vital. The consequences of selecting items for tests based on the relative achievement level of the examinees and the difficulty of the item should be explored in the same context as the construct and operational definition approaches were studied, with item universes being administered to representative samples and comparison between true and estimated true observations.

The lack of suitable test data has made empirical studies of rivaling approaches to testing impossible. The defining of finite item universes, coupled with the administration of these universes to students in instructional settings, offers many opportunities to examine the consequences of various test theories and respective strategies on test results. Future studies will be directed at making empirical comparisons among these theoretical approaches across a wide range of problems.

REFERENCE NOTES

1. Haladyna, T. M. & Roid, G. The quality of domain-referenced test items. A paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1976.
2. Helmstadter, G. C. Comparison of traditional item analysis selection procedures with those recommended for tests designed to measure achievement following performance-oriented instruction. A paper presented at the annual meeting of the American Psychological Association, Hawaii, 1972.
3. Kosecoff, J. B. & Klein, S. P. Instructional sensitivity statistics appropriate for objective-based test items. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.
4. Cox, R. C. & Vargas, J. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. A paper presented at the annual meeting of the American Educational Research Association, Chicago, 1966.
5. Helmstadter, G. C. A comparison of Bayesian and traditional indexes of test item effectiveness. A paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

REFERENCES

- Bloom, B. S. Learning for mastery. Evaluation Comment, 1968, 1, 1-12
- Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.
- Crehan, K. D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11, 255-262.
- Cronbach, L. J. & Meehl, P. E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.
- Glaser, R. Adapting the elementary school curriculum to individual performance. In Proceedings of the invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1967
- Guttman, L. Integration of test design and analysis. In Proceedings of the 1969 invitational conference on testing problems. Princeton, N.J.: Educational Testing Service, 1969.
- Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.
- Hambleton, R. K., Hutton, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objective-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.
- Hively, W., Patterson, H. L. & Page, S. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.
- Keller, F. S. Goodbye, teacher.... Journal of Applied Behavior Analysis, 1968, 1, 79-89.
- Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.
- Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.
- Millman, J. Sampling plans for domain-referenced tests. Educational Technology, 1974(a), 14, 17-21.
- Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.) Evaluation in education: Current applications. San Francisco: McCutchan, 1974(b).
- Nunnally, J. C. Psychometric theory. New York: McGraw-Hill, 1967.
- Osborn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

REFERENCES, continued

- Popham, W. J. & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.
- Roid, G. H. & Haladyna, T. M. A comparison of objective-based and Bormuth item-writing techniques. Educational and Psychological Measurement, (in press).
- Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-148.